Discovering and validating biological hypotheses from coherent patterns in functional genomics data

Marcin P. Joachimiak

SIM2008

August 12th 2008

2008 SIM

Environmental Microbiology Sessions
Integrated omics in systems biology: The new frontier for environmental biotechnology
Convener: Terry C. Hazen, Univ. of California, Berkeley, CA

Original: “Transcriptomics and Bioinformatics”

“Discovering and validating biological hypotheses from coherent patterns in functional genomics data”

Marcin P. Joachimiak

Abstract

The area of transcriptomics analysis is among the more established in computational biology, having evolved in both technology and experimental design. Transcriptomics has a strong impetus to develop sophisticated computational methods due to the large amounts of available whole-genome datasets for many species and because of powerful applications in regulatory network reconstruction as well as elucidation and modeling of cellular transcriptional responses. While gene expression microarray data can be noisy and comparisons across experiments challenging, there are a number of sophisticated methods that aid in arriving at statistically and biologically significant conclusions. As such, computational transcriptomics analysis can provide guidance for analysis of results from newer experimental technologies. More recently, search methods have been developed to identify modules of genes, which exhibit coherent expression patterns in only a subset of experimental conditions. The latest advances in these methods allow to integrate multiple data types and datasets, both experimental and computational, within a single statistical framework accounting for data confidence and relevance to specific biological questions. Such frameworks provide a unified environment for the exploration of specific biological hypothesis and for the discovery of coherent data patterns along with the evidence supporting them.

ESPP2 is part of the Virtual Institute for Microbial Stress and Survival supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program:GTL through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

Large scale biology: towards model cells and organisms

Focused on rapidly inferring as much as possible about a cell or organism:

- its physiology,
- the networks that control its behavior,
- and how the resultant phenotypes allow them to survive in diverse and uncertain environments.

Inference based on collaborative, high throughput experiments and comparative analysis.

Yeast community
Environment Stress Pathway Project (ESPP)
Virtual Institute for Microbial Stress and Survival (VIMSS)
Protein Complex Analysis Project (PCAP)

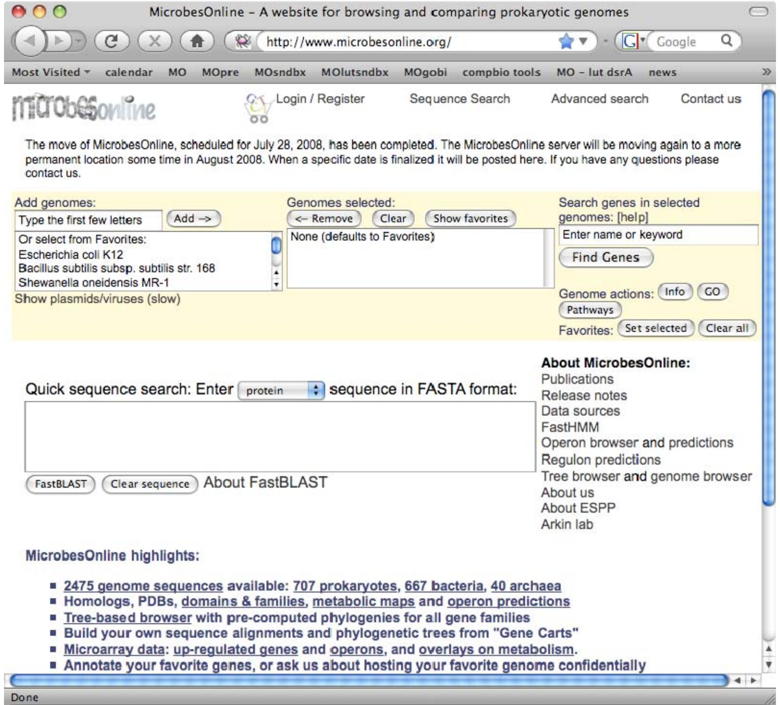
Data from large scale biology projects

- Gene
  - Function
  - Cis-regulatory site
  - Phenotype/fitness (upon gene knockout)
  - Phylogenetic distribution
- Molecular species
  - Relative or absolute level
    - Transcriptomics
    - Proteomics
    - Metabolomics
  - Interaction with other molecular species
    - Protein-protein interactions
    - Transcription factor-DNA binding

Bringing it all together physically and contextually

- Each individual experiment has merit and can be interpreted in isolation.
- However, to achieve the large scale goals it is necessary to integrate datasets and gain multiple contexts.

www.microbesonline.org



The move of MicrobesOnline, scheduled for July 28, 2008, has been completed. The MicrobesOnline server will be moving again to a more permanent location some time in August 2008. When a specific date is finalized it will be posted here. If you have any questions please contact us.

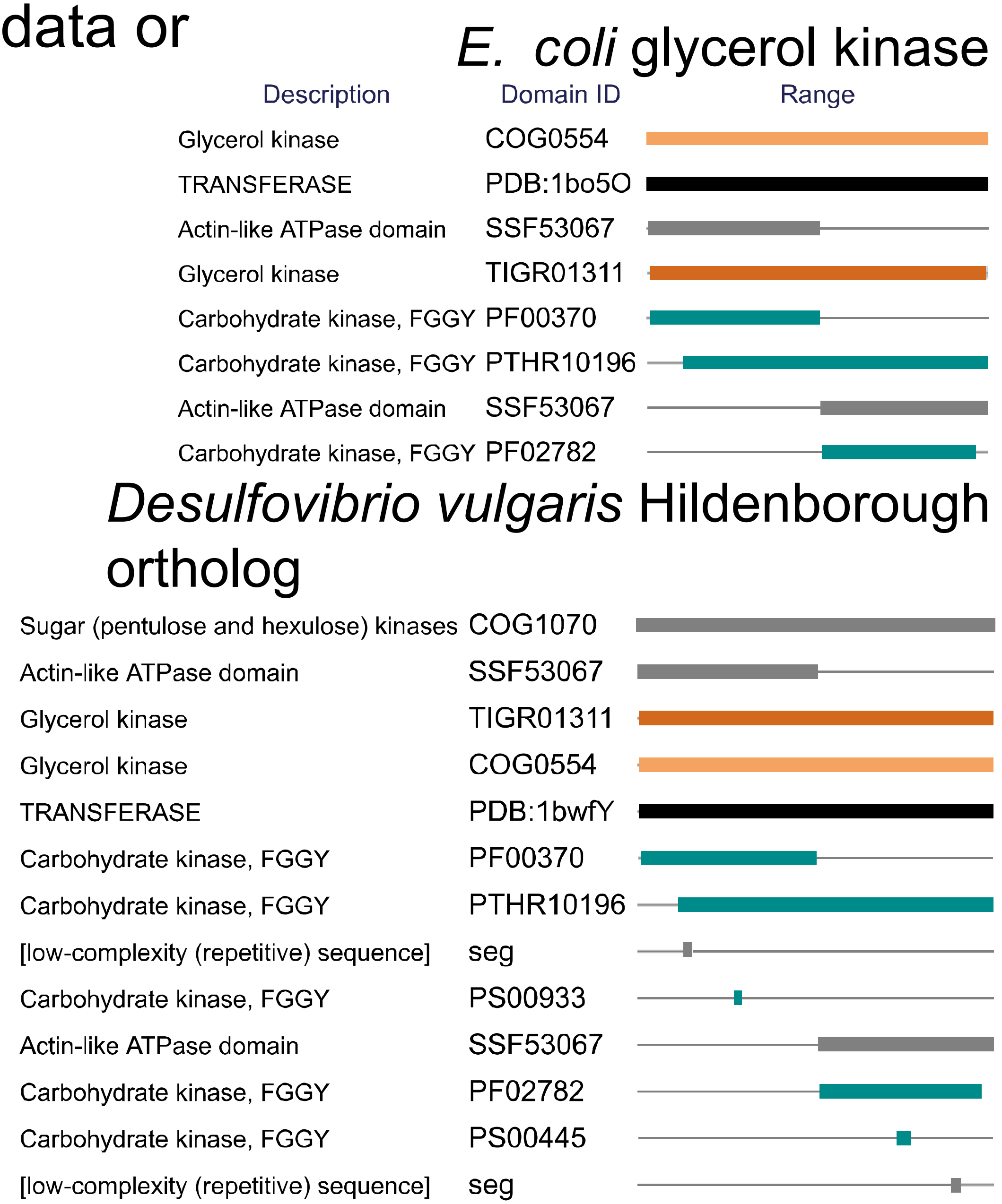
MicrobesOnline highlights:

- 2475 genome sequences available: 707 prokaryotes, 667 bacteria, 40 archaea
- Homologs, PDBs, domains & families, metabolic maps and operon predictions
- Tree-based browser with pre-computed phylogenies for all gene families
- Build your own sequence alignments and phylogenetic trees from "Gene Carts"
- Microarray data: up-regulated genes and operons, and overlays on metabolism.
- Annotate your favorite genes, or ask us about hosting your favorite genome confidentially

Gene function inference

In absence of direct experimental data or characterized close orthologs ...

- Sequence similarity
- Domain architecture
- Functional residues
- Genome location context and presence of expected associates
- 3D structure modeling



E. coli glycerol kinase

Description	Domain ID	Range
Glycerol kinase	COG0554	
TRANSFERASE	PDB:1bo5O	
Actin-like ATPase domain	SSF53067	
Glycerol kinase	TIGR01311	
Carbohydrate kinase, FGGY	PF00370	
Carbohydrate kinase, FGGY	PTHR10196	
Actin-like ATPase domain	SSF53067	
Carbohydrate kinase, FGGY	PF02782	

Desulfovibrio vulgaris Hildenborough ortholog

Sugar (pentulose and hexulose) kinases	COG1070
Actin-like ATPase domain	SSF53067
Glycerol kinase	TIGR01311
Glycerol kinase	COG0554
TRANSFERASE	PDB:1bwfY
Carbohydrate kinase, FGGY	PF00370
Carbohydrate kinase, FGGY	PTHR10196
[low-complexity (repetitive) sequence]	seg
Carbohydrate kinase, FGGY	PS00933
Actin-like ATPase domain	SSF53067
Carbohydrate kinase, FGGY	PF02782
Carbohydrate kinase, FGGY	PS00445
[low-complexity (repetitive) sequence]	seg

Legend

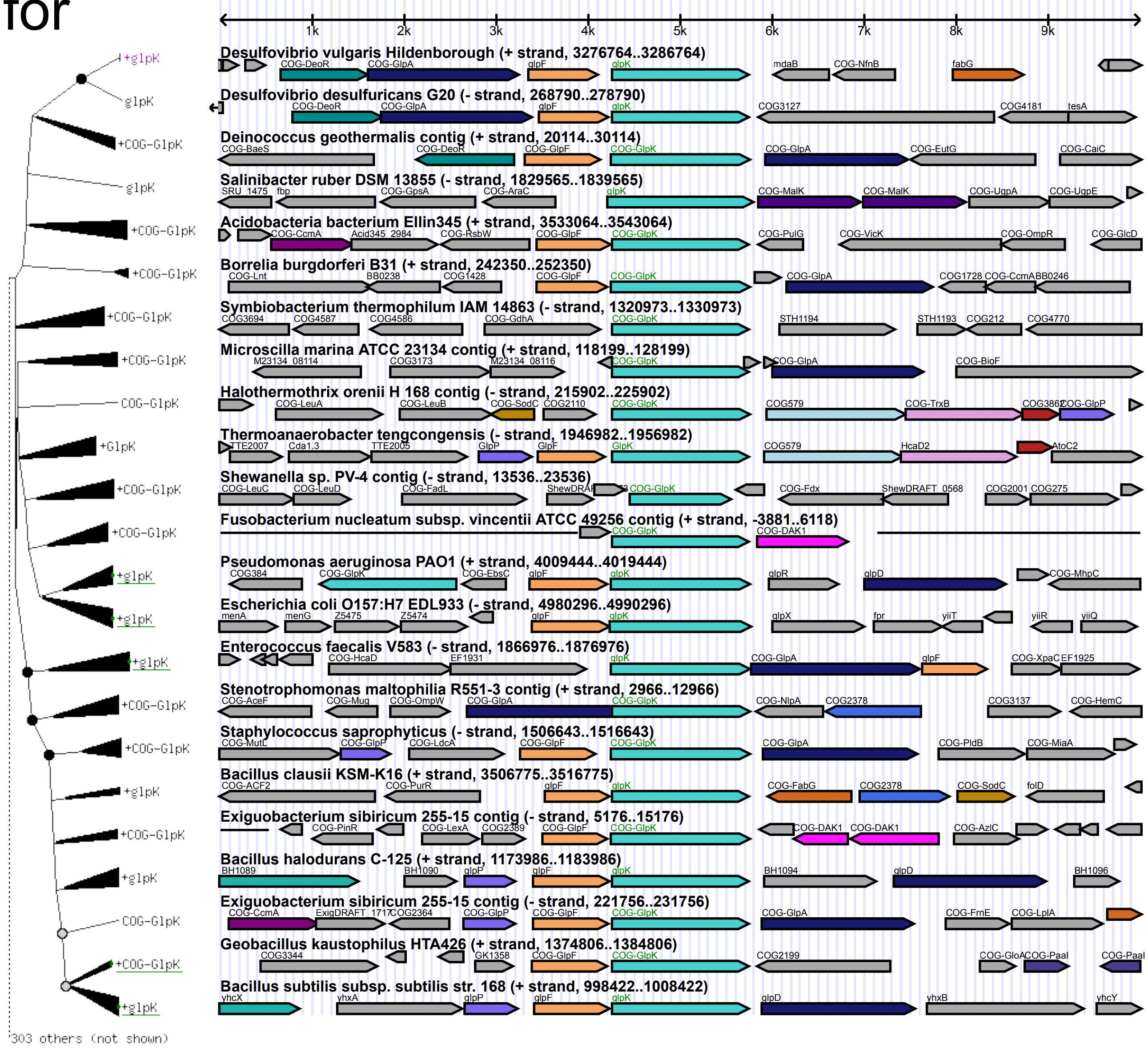
InterPro: IPR000577: 🟩 IPR005999: 🟧

Best COG: 🟧 No IPR/Other COGs: ▩ PDBs: ⬛

MicrobesOnline comparative tools

Tree and genome browser for 3,699,361 proteins in 457,623 families from 1076 microbal genomes

(includes multiple family assignments)


+glpK
glpK
+COG-GlpK
glpK
+COG-GlpK
+COG-GlpK
+COG-GlpK
+COG-GlpK
COG-GlpK
+GlpK
+COG-GlpK
+COG-GlpK
+glpK
+glpK
+glpK
+COG-GlpK
+COG-GlpK
+glpK
+COG-GlpK
+glpK
COG-GlpK
+COG-GlpK
+glpK
303 others (not shown)
1k
2k
3k
4k
5k
6k
7k
8k
9k
Desulfovibrio vulgaris Hildenborough (+ strand, 3276764..3286764)
Desulfovibrio desulfuricans G20 (- strand, 268790..278790)
Deinococcus geothermalis contig (+ strand, 20114..30114)
Salinibacter ruber DSM 13855 (- strand, 1829565..1839565)
Acidobacteria bacterium Ellin345 (+ strand, 3533064..3543064)
Borrelia burgdorferi B31 (+ strand, 242350..252350)
Symbiobacterium thermophilum IAM 14863 (- strand, 1320973..1330973)
Microscilla marina ATCC 23134 contig (+ strand, 118199..128199)
Halothermothrix orenii H 168 contig (- strand, 215902..225902)
Thermoanaerobacter tengcongensis (- strand, 1946982..1956982)
Shewanella sp. PV-4 contig (- strand, 13536..23536)
Fusobacterium nucleatum subsp. vincentii ATCC 49256 contig (+ strand, -3881..6118)
Pseudomonas aeruginosa PAO1 (+ strand, 4009444..4019444)
Escherichia coli O157:H7 EDL933 (- strand, 4980296..4990296)
Enterococcus faecalis V583 (- strand, 1866976..1876976)
Stenotrophomonas maltophilia R551-3 contig (+ strand, 2966..12966)
Staphylococcus saprophyticus (- strand, 1506643..1516643)
Bacillus clausii KSM-K16 (+ strand, 3506775..3516775)
Exiguobacterium sibiricum 255-15 contig (- strand, 5176..15176)
Bacillus halodurans C-125 (+ strand, 1173986..1183986)
Exiguobacterium sibiricum 255-15 contig (- strand, 221756..231756)
Geobacillus kaustophilus HTA426 (+ strand, 1374806..1384806)
Bacillus subtilis subsp. subtilis str. 168 (+ strand, 998422..1008422)

Counts for upcoming release of www.microbesonline.org

MicrobesOnline and RegTransBase

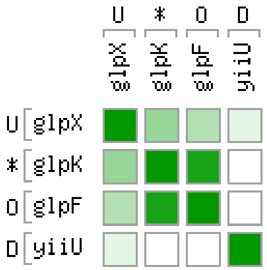
- RegTransBase provides information on microbial transcription factor binding sites and sequence motifs based on expert curation and literature.

Articles Curated:	4,445
Experiments:	10,216
Organisms:	180
Genes:	17,346
Sites:	8,833
Regulators:	971
Effectors:	794

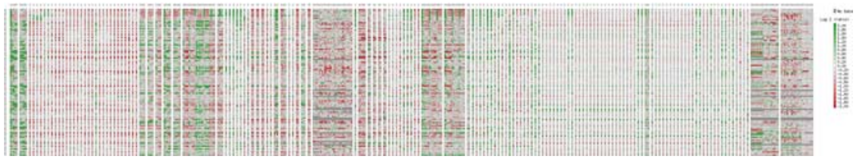
- The two databases and websites are interlinked with ongoing development efforts for new functionalities.

MicrobesOnline data analysis

Currently focused on gene expression data but generalizable to many data types.

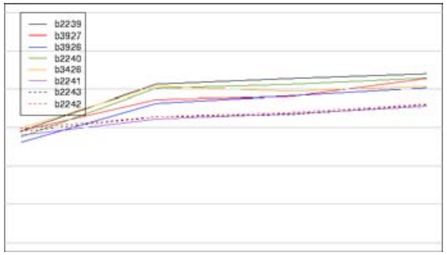


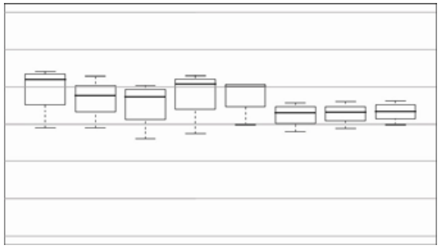
- Gene-gene and experiment-experiment correlations *



- Gene expression profile searches

  (functional profiling)

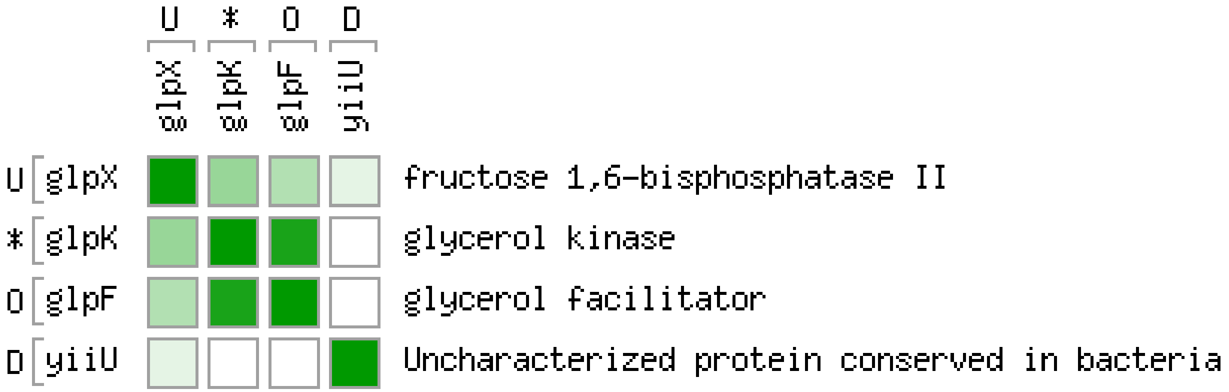


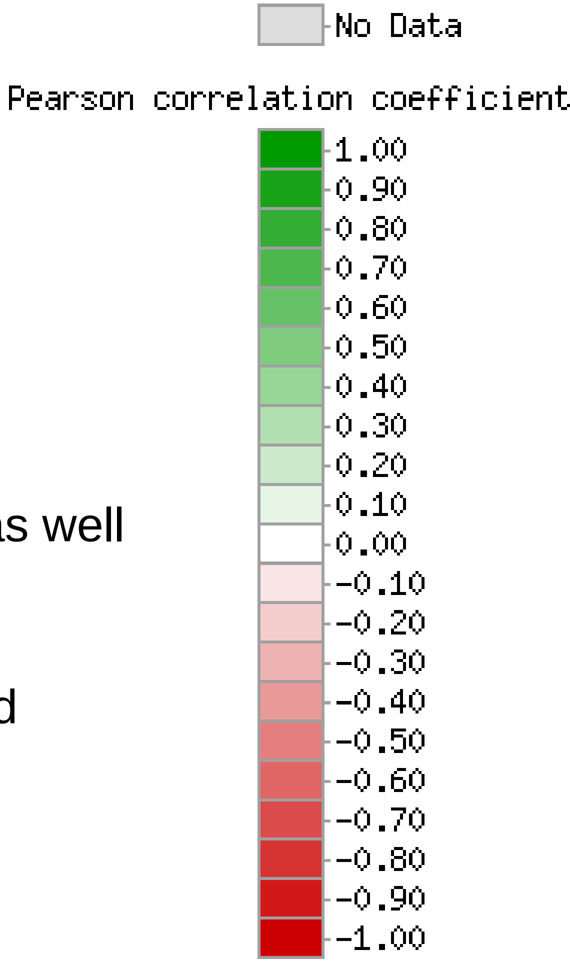


- Line and box plots allowing to subset on genes and experiments *

*Upcoming release of www.microbesonline.org

Analysis:
gene-gene expression correlations





Summarizes similarity in gene expression for a set of genes across a set of experiments.

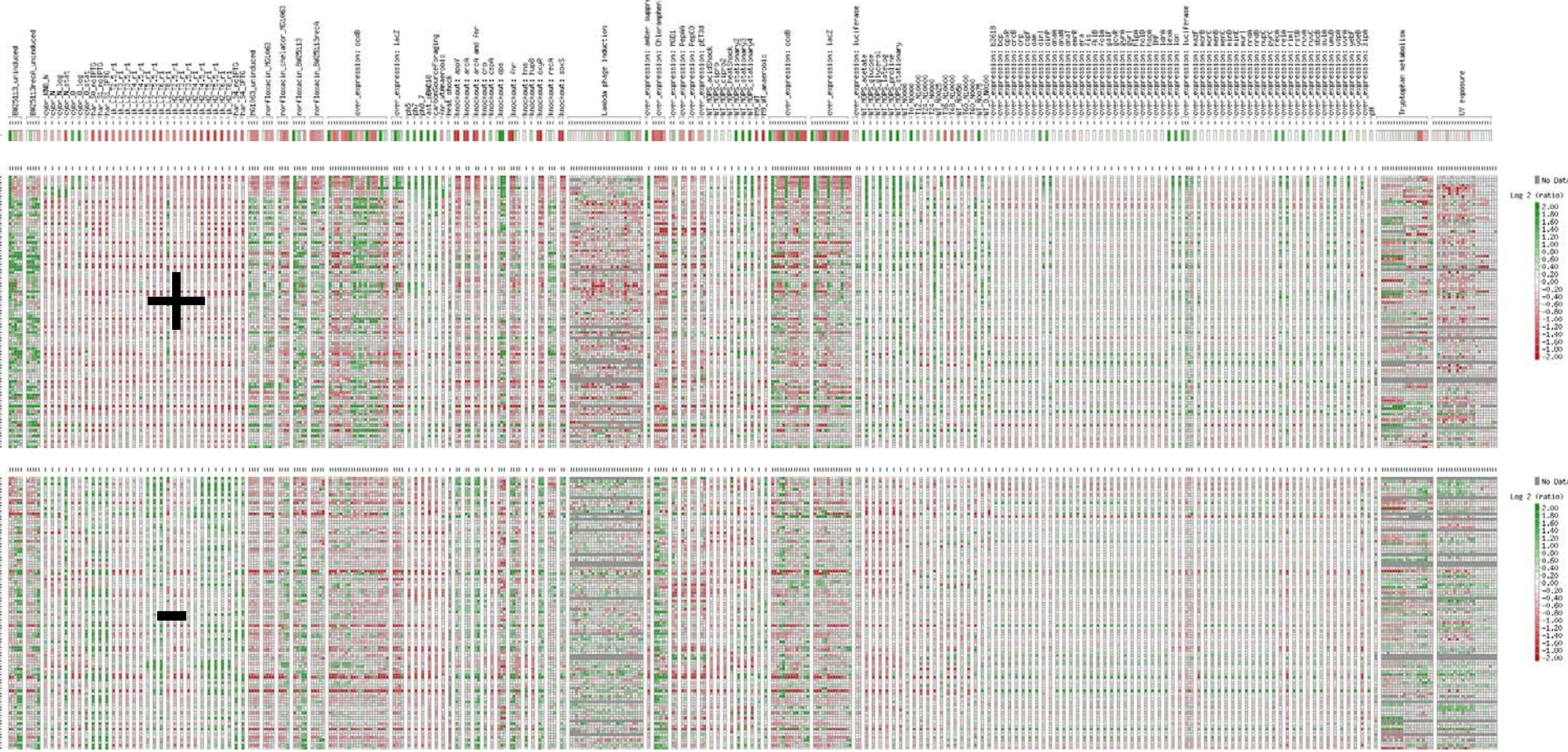
Genes flanking an operon provide a context for significance as well as operon assignments.

Average correlations of randomly sampled genes or permuted gene expression data for the genes of interest can provide statistical significance.

www.microbesonline.org

Gene expression profile searches

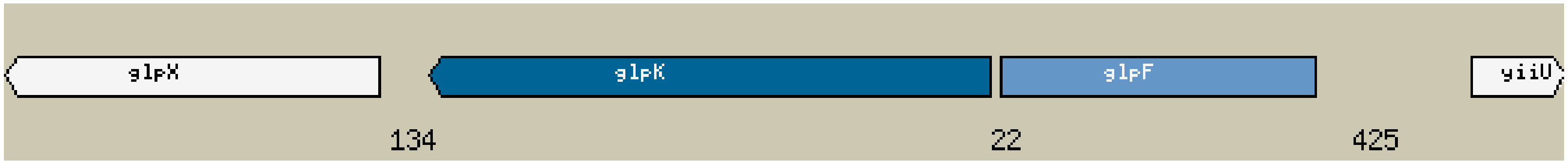
Searches rely on the Pearson correlation coefficient as the similarity measure.
Can identify genes with similar as well as opposite expression patterns.
Computed over many conditions gene co-expression provides expectations
for future experiments and has applications in engineering and design.

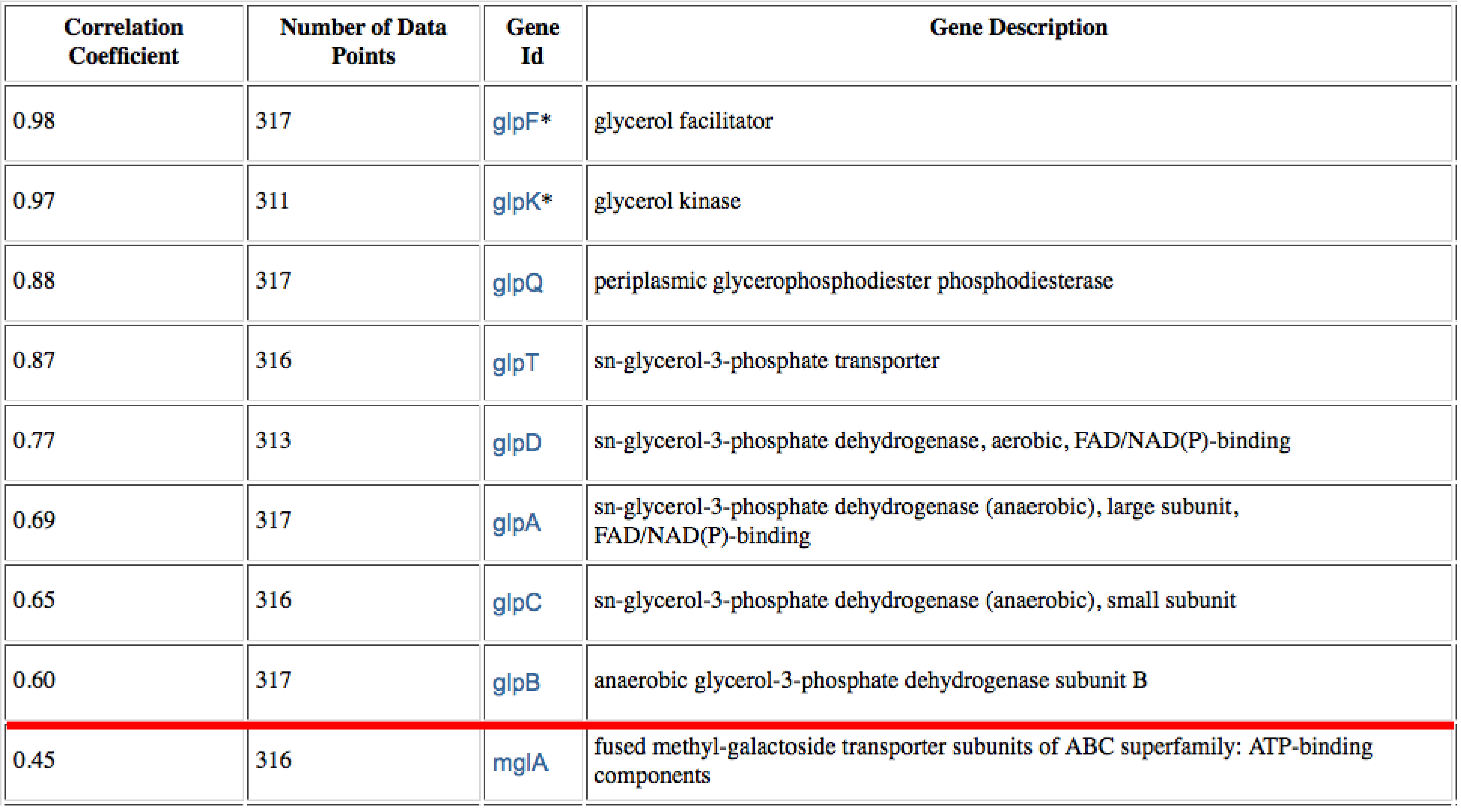


www.microbesonline.org

Identifying candidate genes with gene expression profile searches

This example query profile was based on the mean expression of two *E. coli* genes from a single operon. A biologically motivated cutoff appeared based on a group of clearly functionally related genes from the glycerophospholipid metabolism pathway.



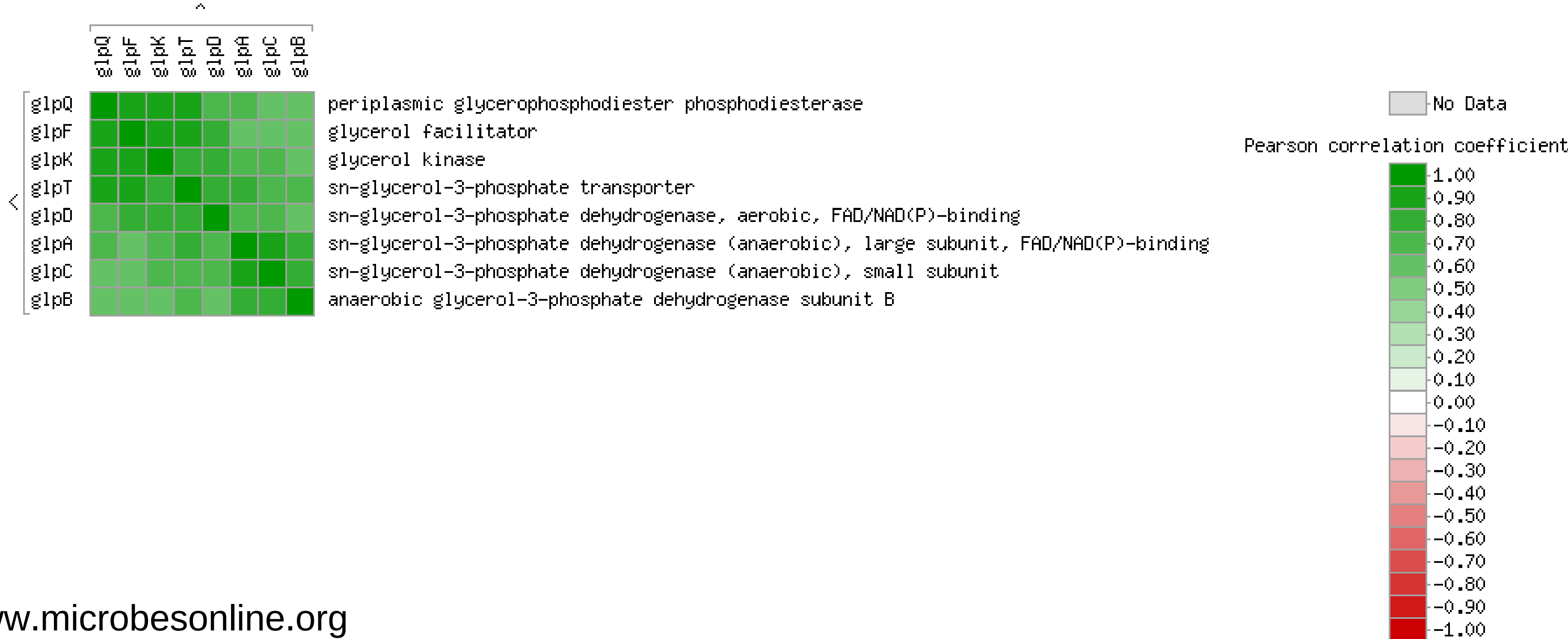


Correlation Coefficient	Number of Data Points	Gene Id	Gene Description
0.98	317	glpF*	glycerol facilitator
0.97	311	glpK*	glycerol kinase
0.88	317	glpQ	periplasmic glycerophosphodiester phosphodiesterase
0.87	316	glpT	sn-glycerol-3-phosphate transporter
0.77	313	glpD	sn-glycerol-3-phosphate dehydrogenase, aerobic, FAD/NAD(P)-binding
0.69	317	glpA	sn-glycerol-3-phosphate dehydrogenase (anaerobic), large subunit, FAD/NAD(P)-binding
0.65	316	glpC	sn-glycerol-3-phosphate dehydrogenase (anaerobic), small subunit
0.60	317	glpB	anaerobic glycerol-3-phosphate dehydrogenase subunit B
0.45	316	mglA	fused methyl-galactoside transporter subunits of ABC superfamily: ATP-binding components

www.microbesonline.org

Analysis of candidates:
gene-gene expression correlations

Can expand analysis to genome-wide with clustering, but standard clustering is limited and eventually need more sophisticated and computationally intensive bicluster search methods.
For example, need the ability to seed a search with specific genes and experiments.

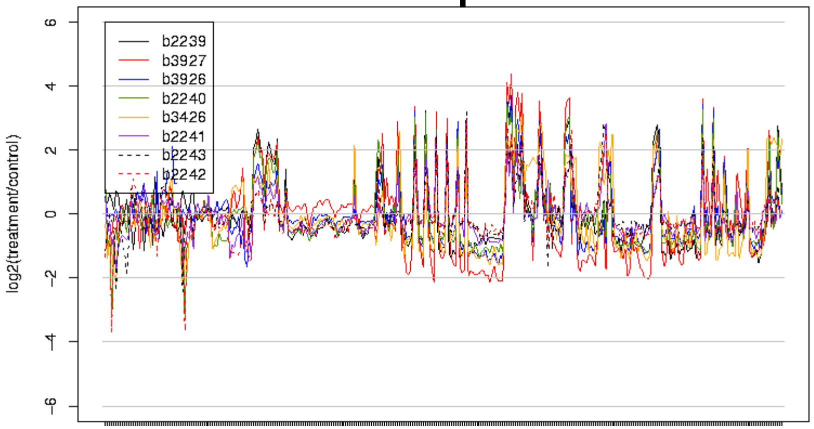


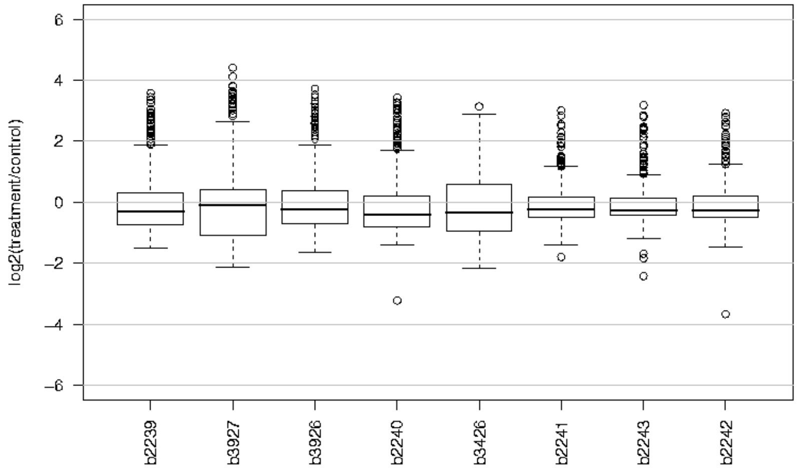
glpQ	periplasmic glycerophosphodiester phosphodiesterase
glpF	glycerol facilitator
glpK	glycerol kinase
glpT	sn-glycerol-3-phosphate transporter
glpD	sn-glycerol-3-phosphate dehydrogenase, aerobic, FAD/NAD(P)-binding
glpA	sn-glycerol-3-phosphate dehydrogenase (anaerobic), large subunit, FAD/NAD(P)-binding
glpC	sn-glycerol-3-phosphate dehydrogenase (anaerobic), small subunit
glpB	anaerobic glycerol-3-phosphate dehydrogenase subunit B

www.microbesonline.org

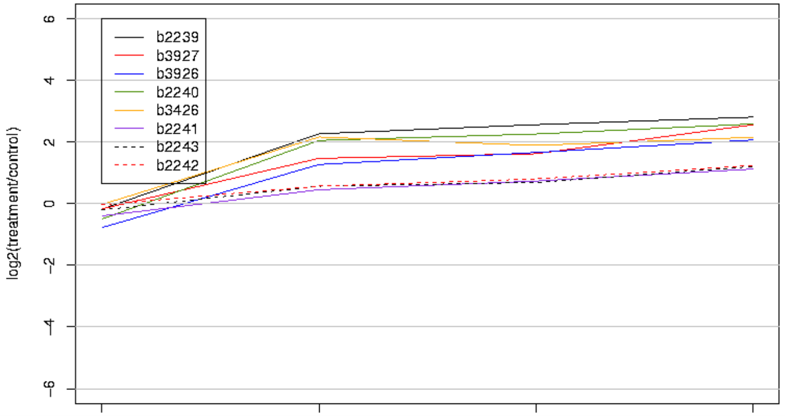
Analysis of candidates: line and box plots

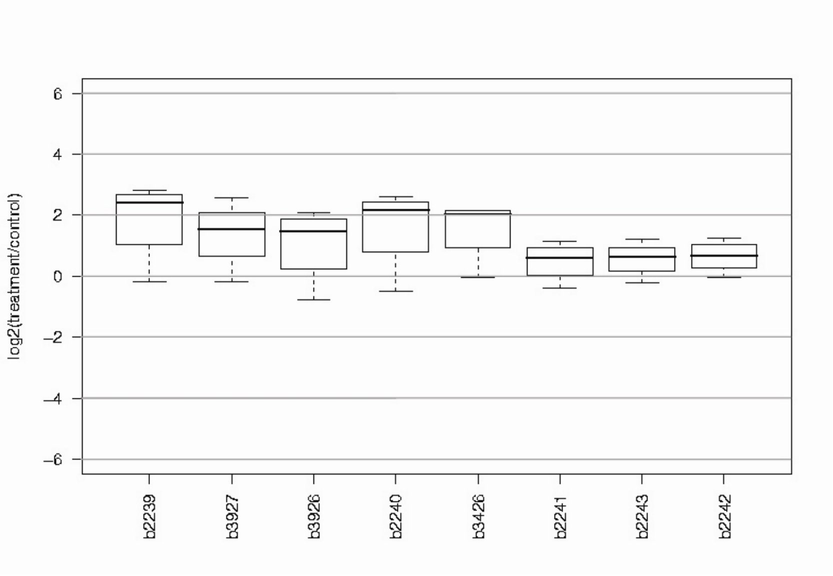
Entire compendium





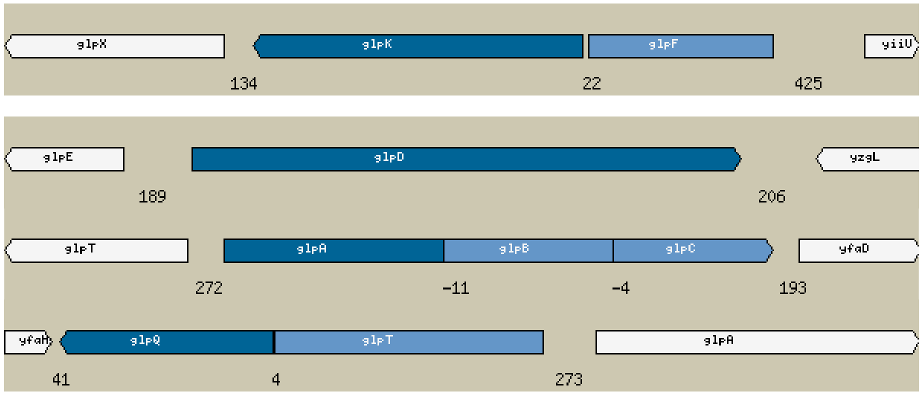
pH 2, 5, 7, 8.7



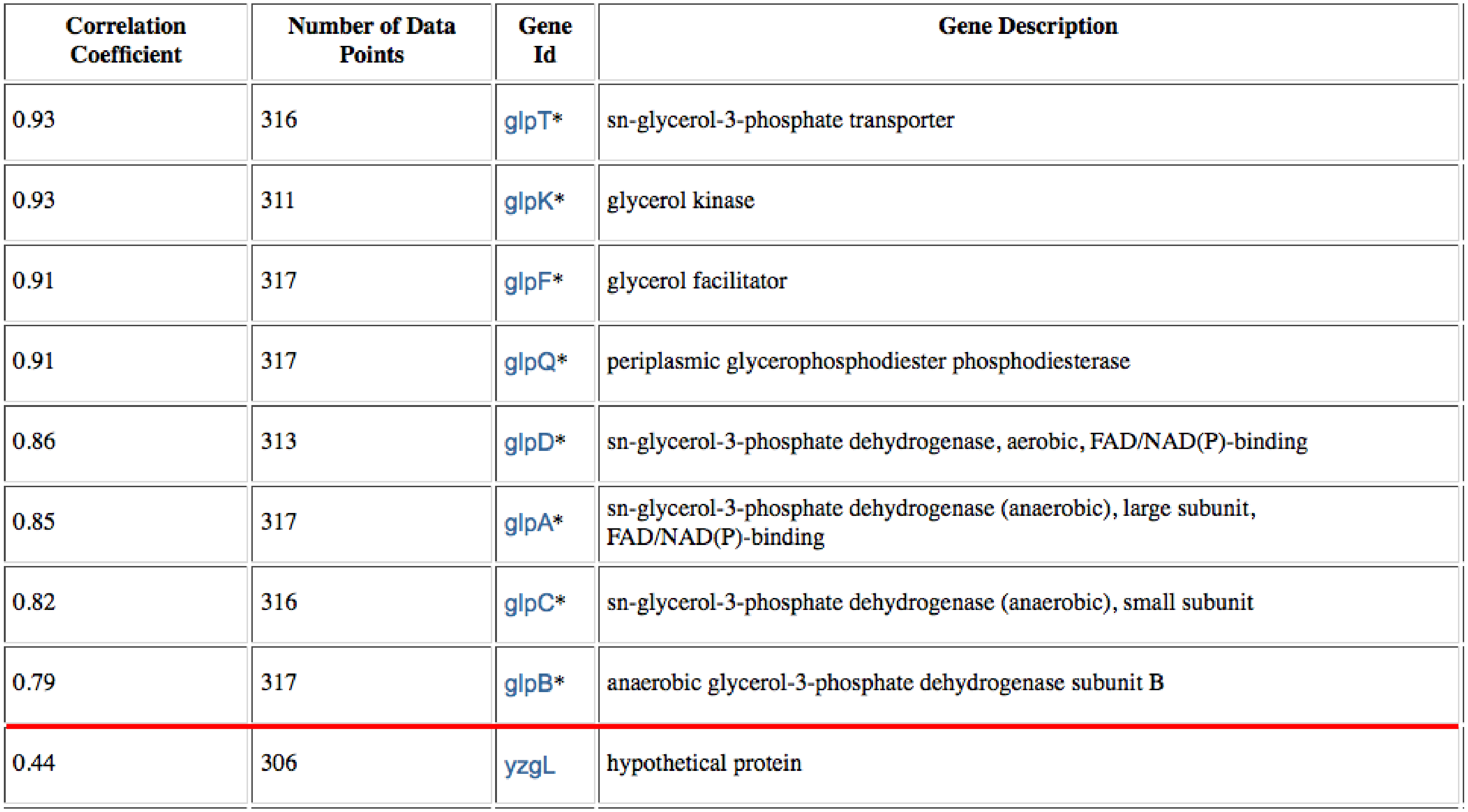


www.microbesonline.org

Analysis of candidates: iterative profile searches



Six additional genes (from 3 different operons) were identified with high expression profile similarity to the query. A natural cutoff was provided by the next best weakly correlated gene. All top hits are members of the glycerophospholipid metabolism pathway.



Correlation Coefficient	Number of Data Points	Gene Id	Gene Description
0.93	316	glpT*	sn-glycerol-3-phosphate transporter
0.93	311	glpK*	glycerol kinase
0.91	317	glpF*	glycerol facilitator
0.91	317	glpQ*	periplasmic glycerophosphodiester phosphodiesterase
0.86	313	glpD*	sn-glycerol-3-phosphate dehydrogenase, aerobic, FAD/NAD(P)-binding
0.85	317	glpA*	sn-glycerol-3-phosphate dehydrogenase (anaerobic), large subunit, FAD/NAD(P)-binding
0.82	316	glpC*	sn-glycerol-3-phosphate dehydrogenase (anaerobic), small subunit
0.79	317	glpB*	anaerobic glycerol-3-phosphate dehydrogenase subunit B
0.44	306	yzgL	hypothetical protein

www.microbesonline.org

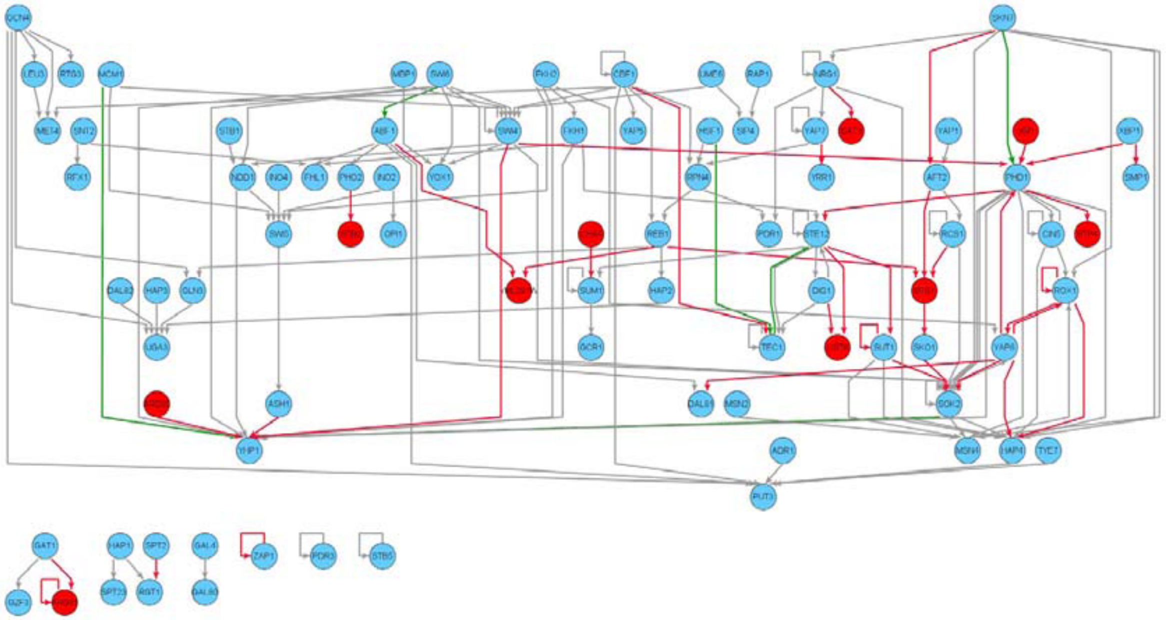
Biological network inference

- Goal is to infer (reverse engineer) the topology of a network (e.g., regulatory, signaling), based on direct, indirect or combined association data.

- Identify data patterns that indicate causal influence.

- Networks serve as the basis for dynamical systems modeling and ultimately prediction of cellular responses.

Nodes = Transcription factors
Edges = regulatory events



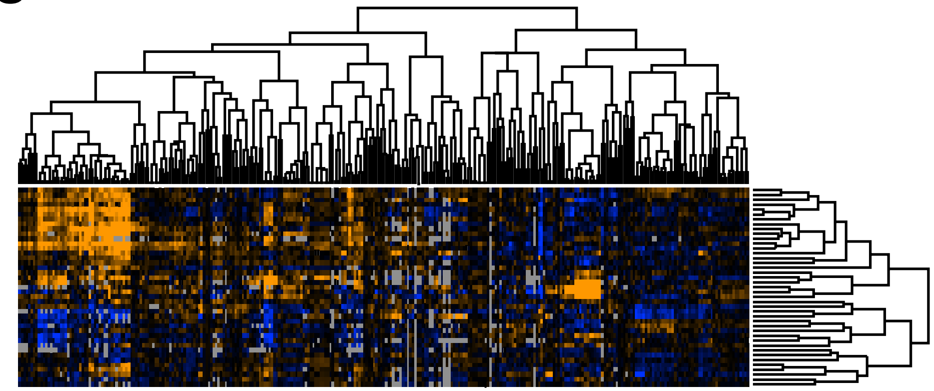
Yeast regulatory network from MacIsaac et al. 2006, based on computational refinement of chromatin immunoprecipitation on chip (ChIP/chip) TF binding site data.

Basic biclustering

A **bicluster** of a data set is a subset of rows that exhibit similar patterns across a subset of columns, or vice versa.

A family of data clustering methods which use a similarity measure (Euclidean distance, Pearson correlation etc.) to compute the input distance matrix for a data clustering algorithm (e.g., hierarchical, k-means). The result are clusters of genes and experiments with similar expression profiles and a 2-D ordering of the data induced by the dendrograms.

Disadvantage: uses all experiments and all genes in comparisons and only 1 bicluster per gene/experiment.



TIGR Multi-experiment viewer, MeV

The next level: bicluster searching and its applications

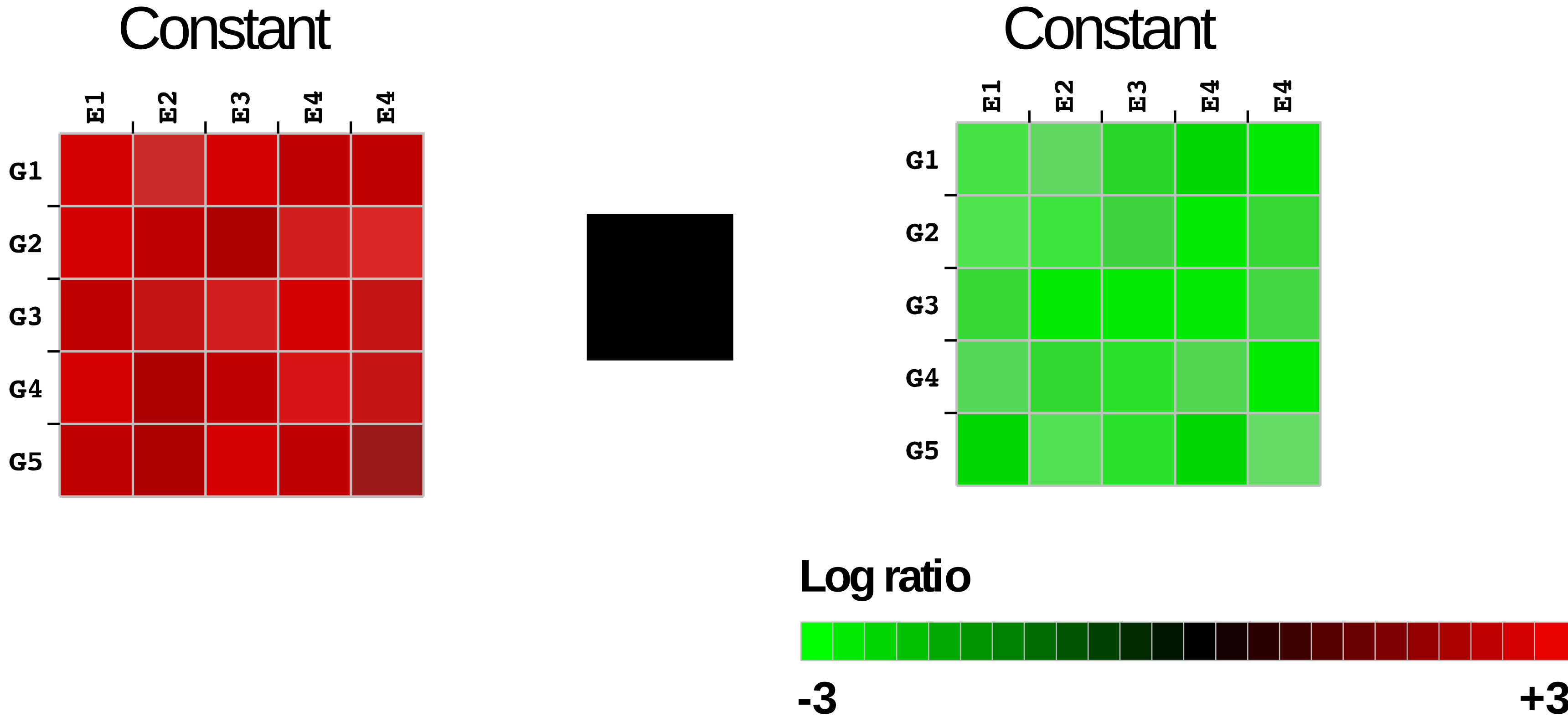
Family of computationally intensive methods, which search a data matrix for potentially overlapping biclusters.

- Discover (overlapping*) sets of genes and experiments exhibiting a (non-binary*) data pattern.
- Test stability of gene+experiment, gene, or experiment sets with respect to a variety* of datasets and statistical criteria.
- Assignment of significance* to patterns in data, removal of noise from results and hypothesis.
- High-throughput experimental data analysis and troubleshooting.

* Increased complexity

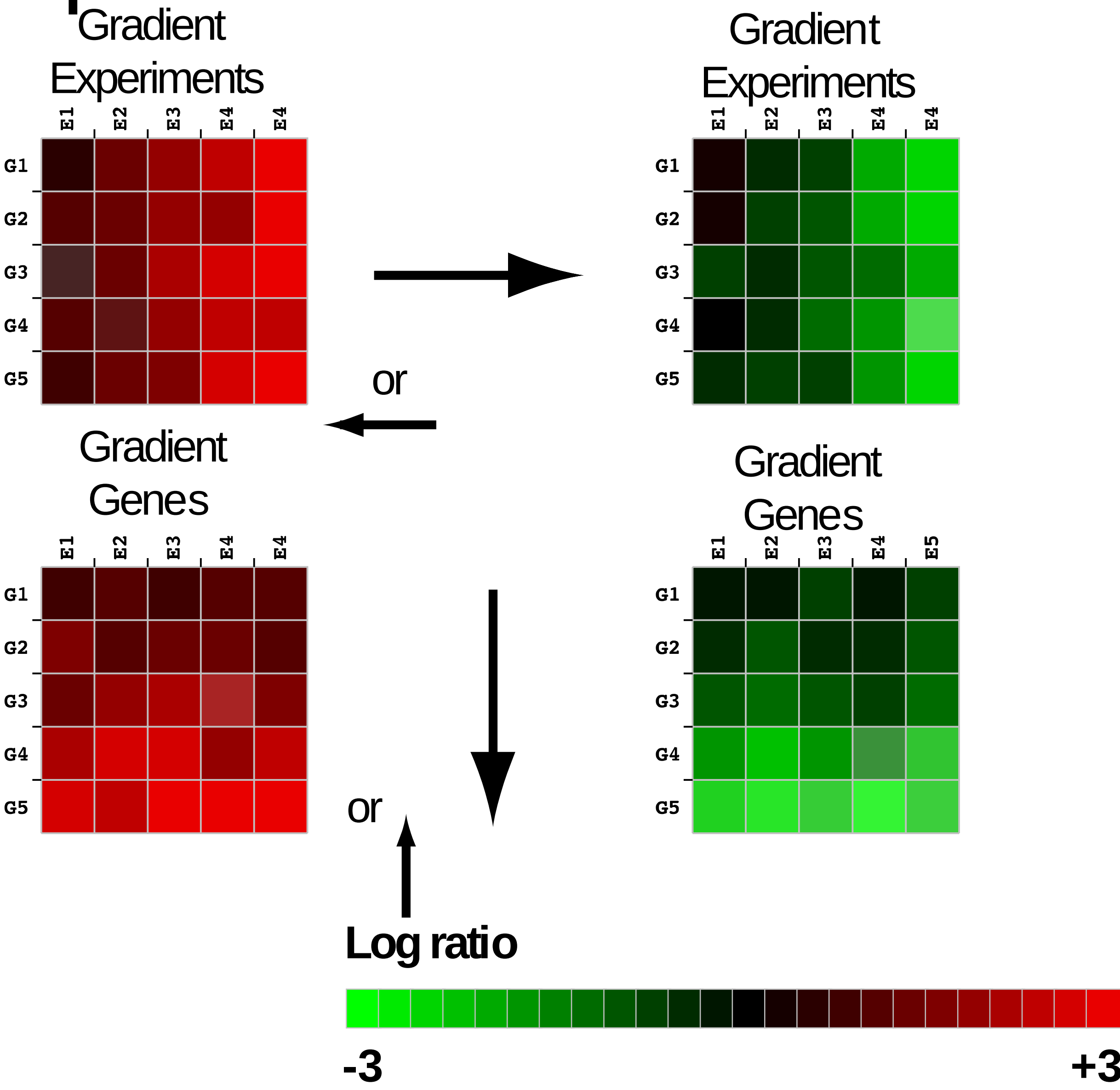
Coherent bicluster patterns: biological mechanism signatures

Uniform differential expression:
Common regulation in the form of transcriptional activation or repression.
E.g., genes in an operon, regulon.



Non-constant coherent bicluster patterns

Increasing/decreasing response in time course or condition gradient.



Gradient Experiments

Gradient Experiments

or

Gradient Genes

Gradient Genes

or

Log ratio

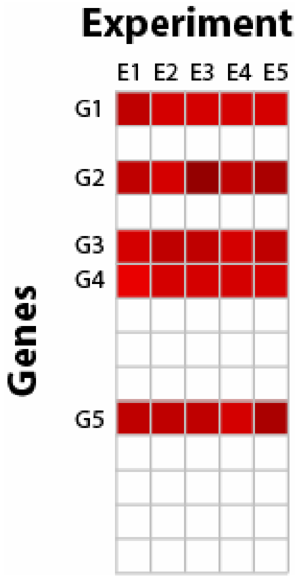
-3

+3

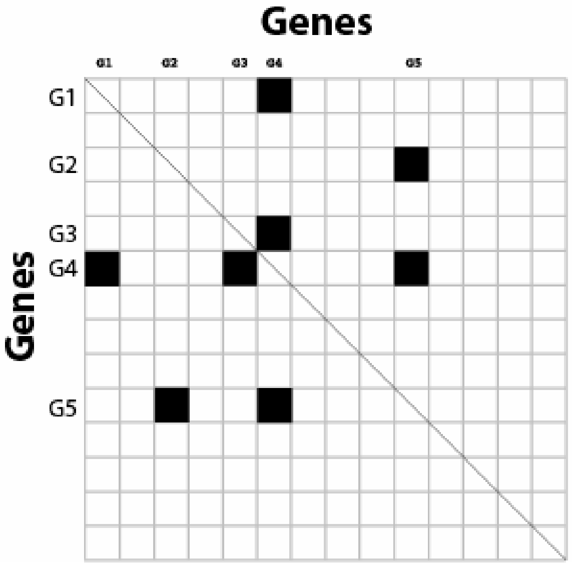
Potentially co-regulated transcripts differing in magnitude of differential expression.

Common functional genomics data types

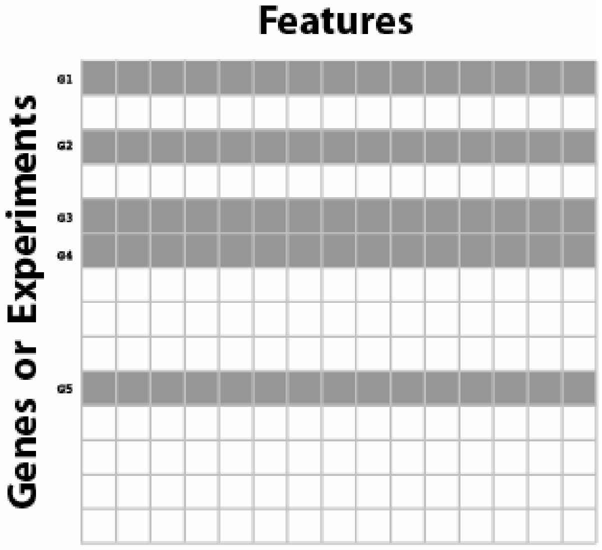
Gene-by-experiment, e.g., gene expression



Gene-by-gene, e.g., protein interaction matrix

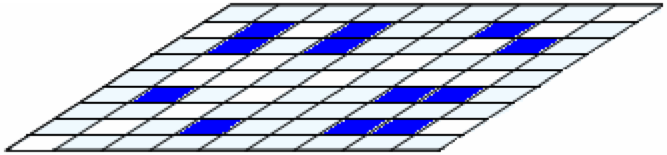


Gene- or experiment- by-features, e.g.:
GO terms
Functional categories
Phylogenetic profiles
Gene, mRNA, and protein features

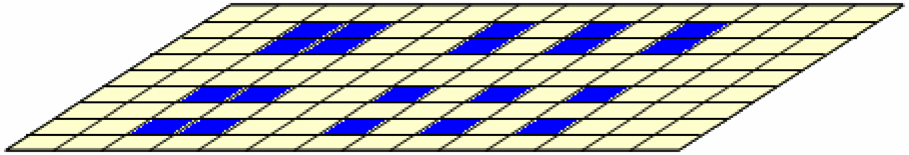


Biclustering as statistical data organization and integration

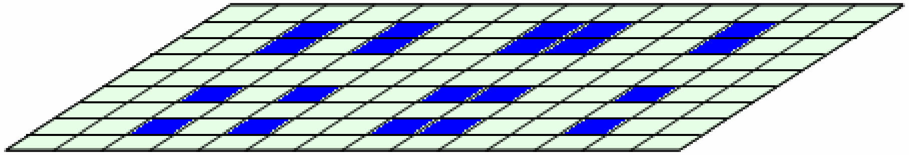
Full criterion = weighted sum of sub-criteria



Protein by Protein



Gene by Experiment



Gene/Experiment by Feature

Manuscript in preparation

Proportion score

$$\mathrm{MSE}(\overline{X}) = \mathrm{E}((\overline{X} - \mu)^2) = \left(\frac{\sigma}{\sqrt{n}}\right)^2$$

Row **Mean Squared Error** (MSE), other correlation and rank criteria

- Significance score calculated from an empirical null distribution created from random draws of all allowed bicluster sizes.
- Probability that a value as extreme or more extreme would occur if bicluster was randomly sampled.

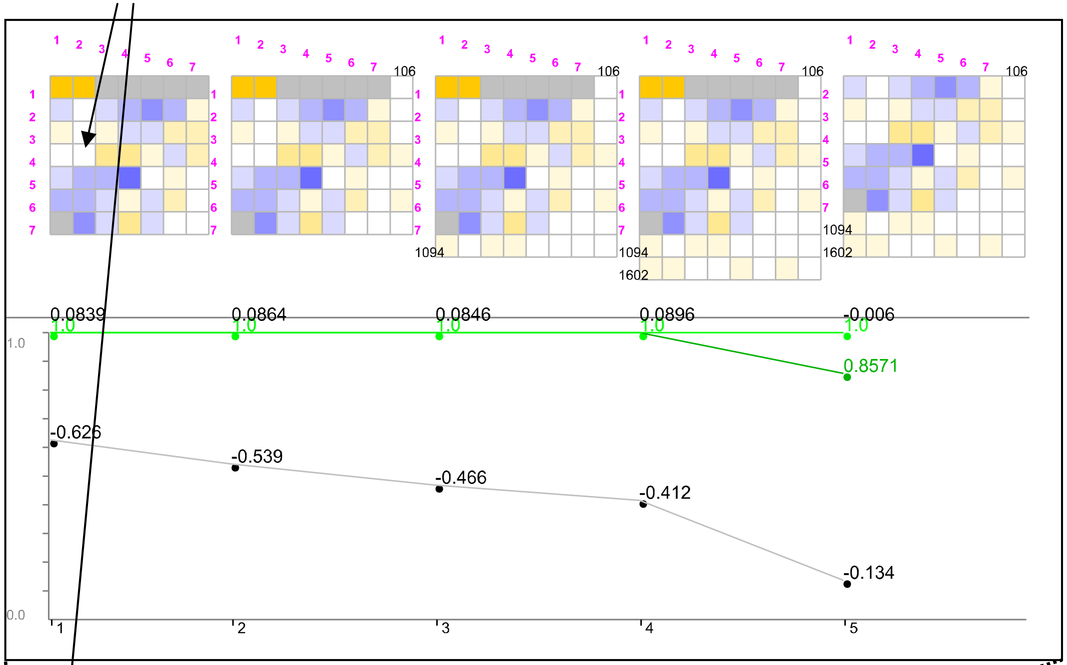
$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Cross-validated R²

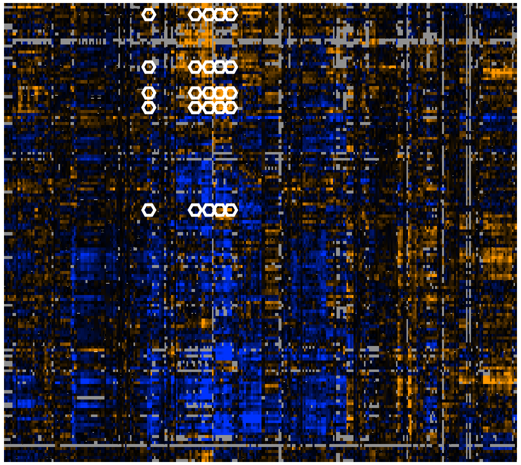
- Calculated using data-adaptive software with polynomial spline fitting.
- Selects subset of features using cross-validation.

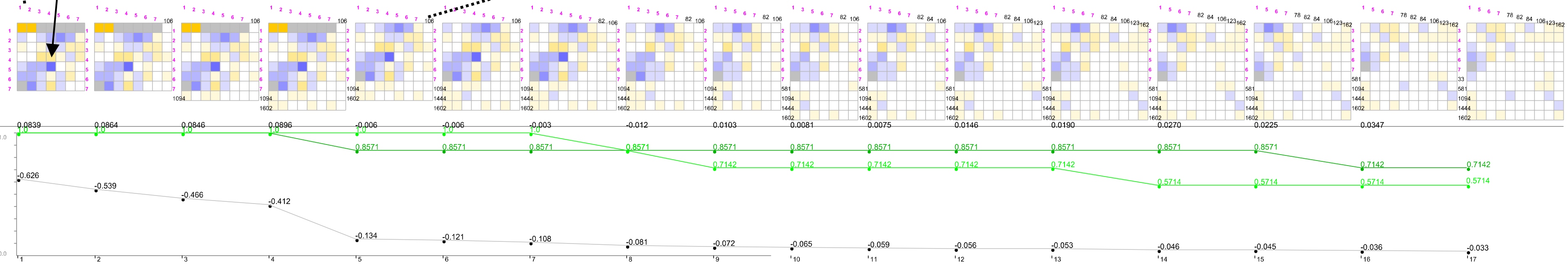
Bicluster search trajectory

Random 7x7 starting block



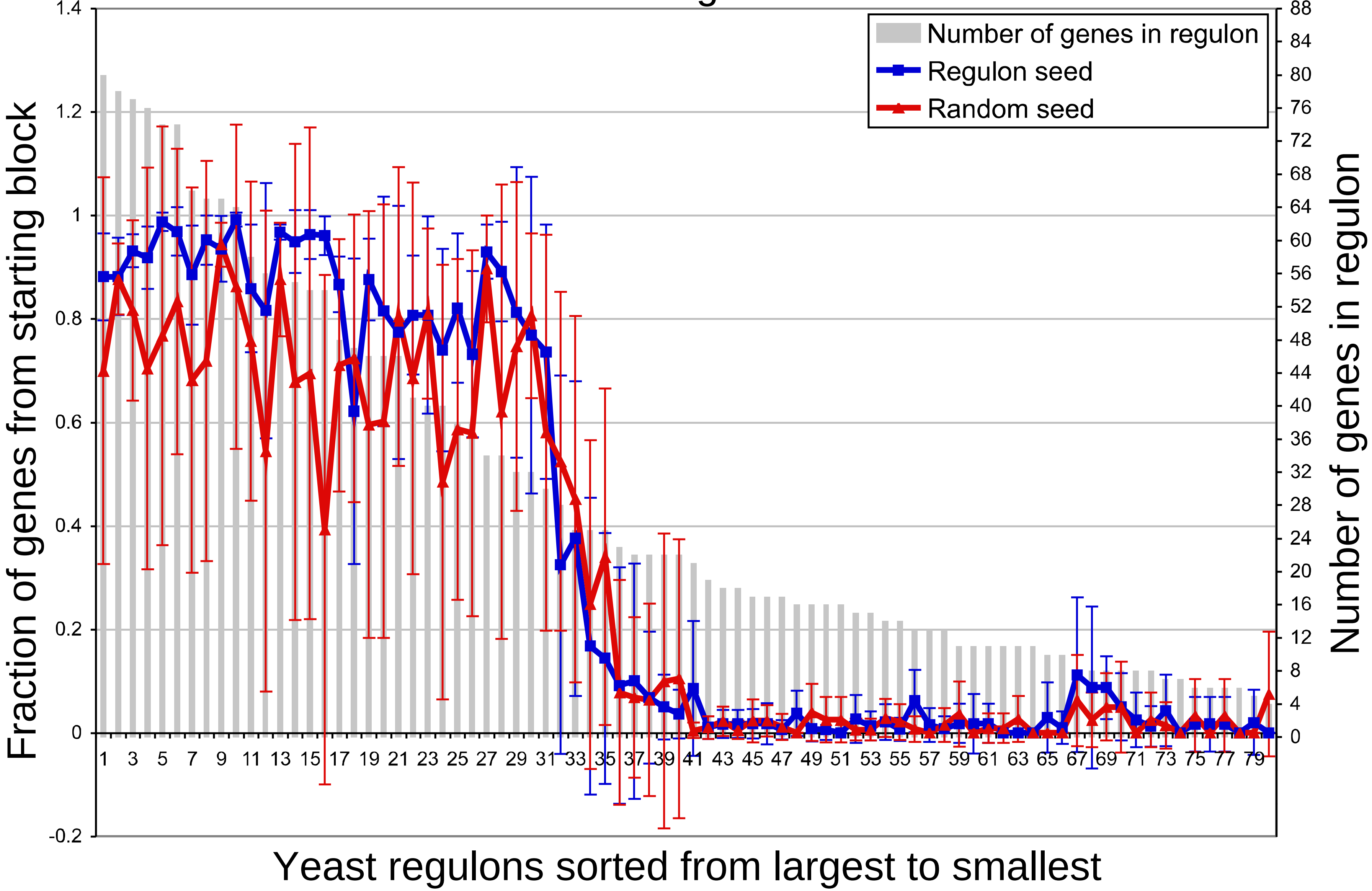
Current bicluster



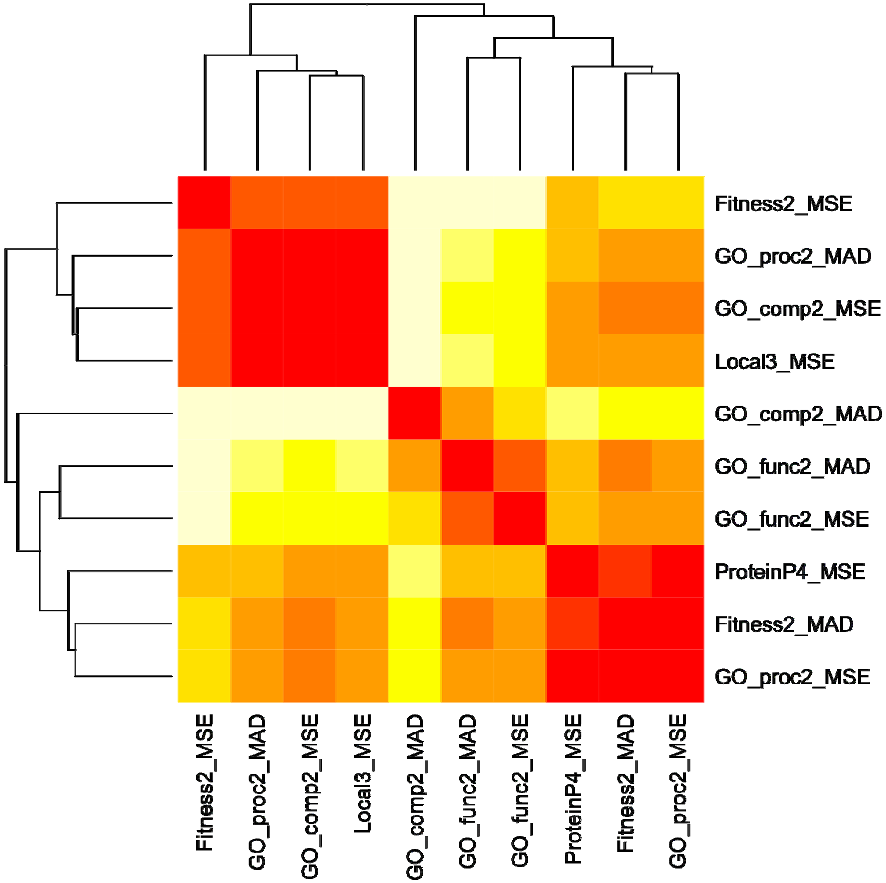


Yeast regulon ‘stability’

Results from novel statistical data fusion algorithm:
10 runs each for 80 *S. cerevisiae* regulons from MacIsaac *et al*. 2006



Comparing features and criteria: bicluster membership



Bicluster **A** and **B** overlap:

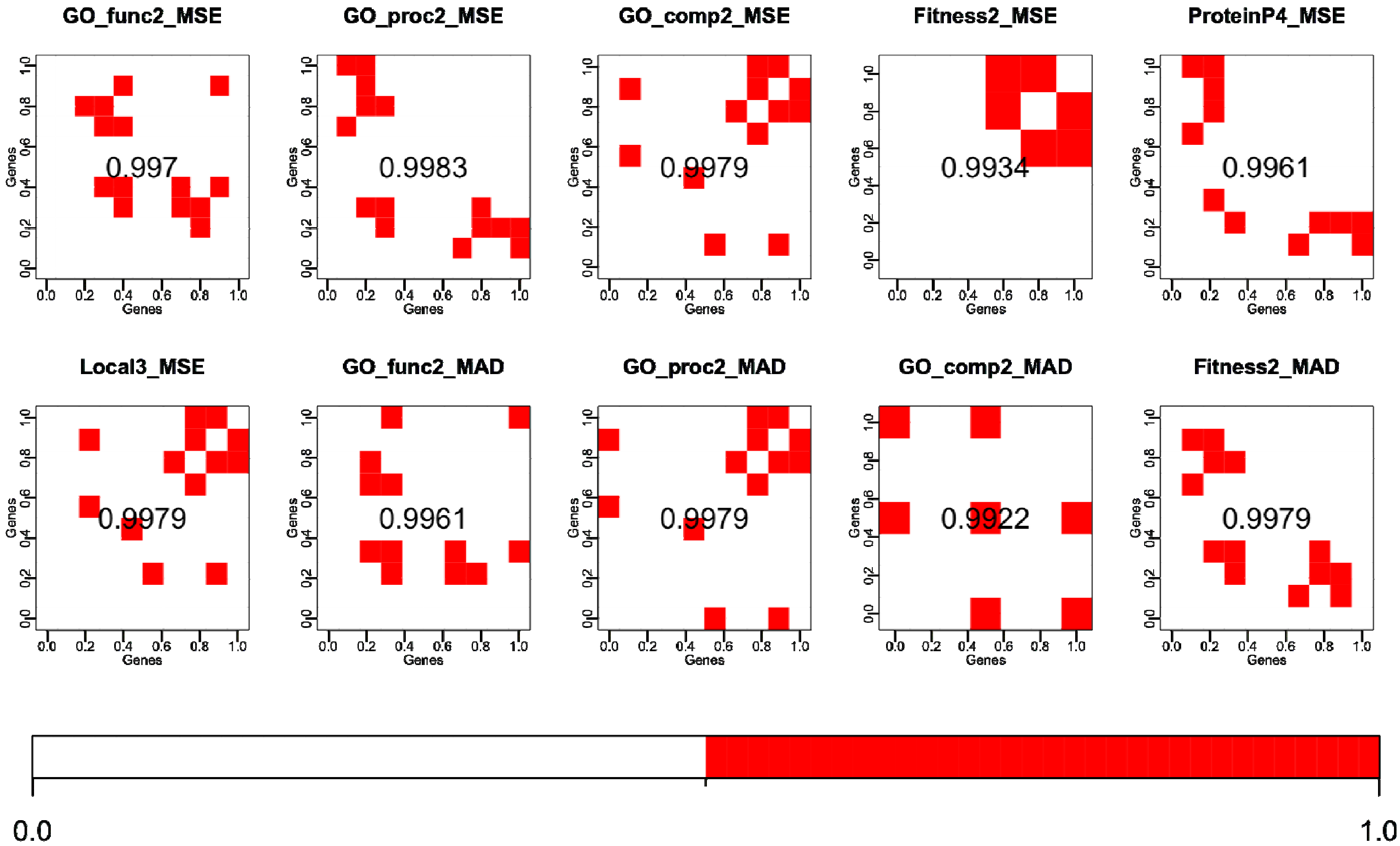
$$\frac{\bigcap \; \text{Genes}_A, \text{Genes}_B}{\sqrt{\text{Genes}_A \text{ x } \text{Genes}_B}}$$

$$\frac{\bigcap \; \text{Experiments}_A, \text{Experiments}_B}{\sqrt{\text{Experiments}_A \text{ x } \text{Experiments}_B}}$$

10 final biclusters

- Single set of runs
- Same starting bicluster
- Varying feature set
- Varying criteria

Protein-protein interaction profiles for final biclusters



GO_func2_MSE

GO_proc2_MSE

GO_comp2_MSE

Fitness2_MSE

ProteinP4_MSE

Local3_MSE

GO_func2_MAD

GO_proc2_MAD

GO_comp2_MAD

Fitness2_MAD

Perspectives

- Next questions
  - How do different features and datasets contribute to known regulon recovery?
  - What are the properties of known regulons?
- Next additions to algorithm
  - Methods
    - Forward selection
    - Post-analysis toolbox
  - Datasets
    - Sequence motifs
    - Pathways and metabolites
- What are the hallmarks of success?
  - Evaluate recovery of known regulons in presence of noise
  - Discovery of novel regulons
  - Dynamical modeling based on predicted regulons

Acknowledgements

Biclustering data fusion algorithm

Adam Arkin
Mark van der Laan
Cathy Tuglus

PCAP

Swapnil Chabra
John-Marc Chandonia

VIMSS & www.microbesonline.org

From left to right:

Dylan Chivian
Paramvir Dehal
Morgan Price
…
Adam Arkin
Keith Keller
Jason Baumohl (not shown)